An Improved Real-time Head Detection Method

Jingyi Yu

North China Electric Power University, Baoding 071000, China.

Jingyiyu111@gmail.com

Abstract

In order to improve the speed and accuracy of head detection in dense scenes, this system proposes an improved method based on YOLO v3. By adjusting the data set and regenerating the anchor box, the mAP value is increased from 19.82% to 38.44%, which significantly improve the detection effect.

Keywords

Head detection, Crowded scenes, Object detection.

1. Introduction

As the world's population continues to grow, there are more and more crowded accidents happened in densely populated public areas. However, the traditional manual detection method has low efficiency and high cost, which is far from meeting the requirements. Therefore, research on head detection in crowded scenes is of great significance.

Many approaches have been used in head detection in the past few decades. At present, most people detection in dense environments adopts the methods of head detection or head and shoulder detection. In 2013, Zhu et al.[1] proposed that human head and shoulder contours are used as input features, and the Adaboost algorithm is used for detection. However, in some scenarios, the angle of surveillance video acquisition is difficult to meet the requirements of the algorithm. Cai et al.[2] proposed an enhanced cascading head detection method based on multi-scale local pattern features in 2014, but this method only relies on the color information of the video and is easily affected by shadows and lighting changes. Gao et al.[3] proposed a local part detection method that combined depth and color information in 2016, but this method requires a depth sensor to assist in obtaining image depth information, which is not conducive to the promotion of the algorithm.

Comprehensive analysis of the problems in the above model, we use deep learning-based target detection technology to achieve head detection, without the need for additional equipment to assist, reducing costs, and has a high recognition rate. In order to meet the requirements of real-time performance and detection accuracy, we improve the YOLO v3 algorithm[4].

2. Method

2.1 YOLO

Our method is based on YOLO v3. The core idea of the YOLO algorithm[5] is to convert the target detection problem into a regression problem for solution. The algorithm can directly extract information from the picture and identify the target object through a single convolution network. The YOLO algorithm divides each image into S × S grids. If the center of the target object falls in a grid, then this grid is responsible for detecting the object. Each grid regression predicts B bounding boxes, and finds the confidence score and category probability of each bounding box. Finally, non-maximum suppression is used to filter out the bounding boxes with lower scores to obtain the target prediction bounding box. The detection speed of the YOLO algorithm is very impressive, which can reach 45FPS, but it has the problems of missing target objects and inaccurate positioning. The YOLO 9000 algorithm[6] made a series of improvements to the YOLO algorithm, and adopted measures such as introducing Batch Normalization, multi-scale training, and dimensional clustering.

YOLO v3 algorithm has made further adaptability improvements on the basis of YOLO 9000, such as using multi-scale recognition, multi-label classification, etc., and it uses DarkNet-53 network as a feature extractor[7]. The YOLO v3 algorithm uses three different scale feature maps for detection, which are 13 × 13, 26 × 26, 52 × 52. The fusion method on the multi-scale feature maps for detection has significantly improved the detection effect of small targets. In addition, the deep network Darknet-53 proposed by the author based on the residual network idea makes the network have better performance through multiple 3 × 3 and 1 × 1 convolutional layer. And on this basis, this deep network can achieve twice the efficiency of resnet-152, making the YOLO v3 algorithm move to the top of the target detection algorithm[8].

2.2 Improvement based on YOLO algorithm

YOLO v3 uses the COCO dataset for training. In the COCO data set, there are many types of target objects, and there is a large gap in the size of different types of objects, as shown in Figure 1 below.

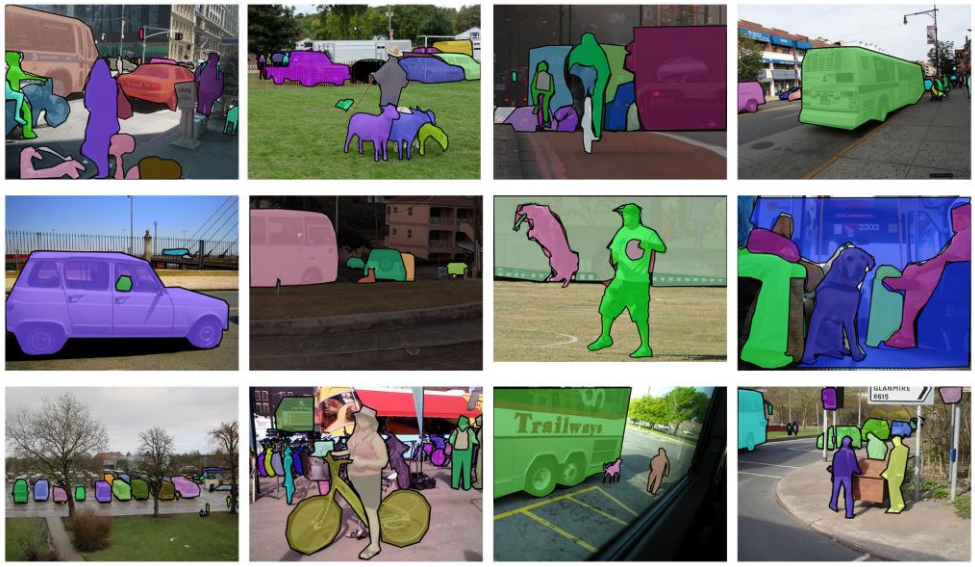

Fig. 1 COCO dataset

In the application scenario of this article, there is only one category of human head, and the proportion of human head in the picture is small and the distribution is dense[9]. If the COCO data set is still used for training, it will affect the detection of small targets such as human heads, so this article uses the Brainwash dense head data set for training, as shown in Figure 2 below.

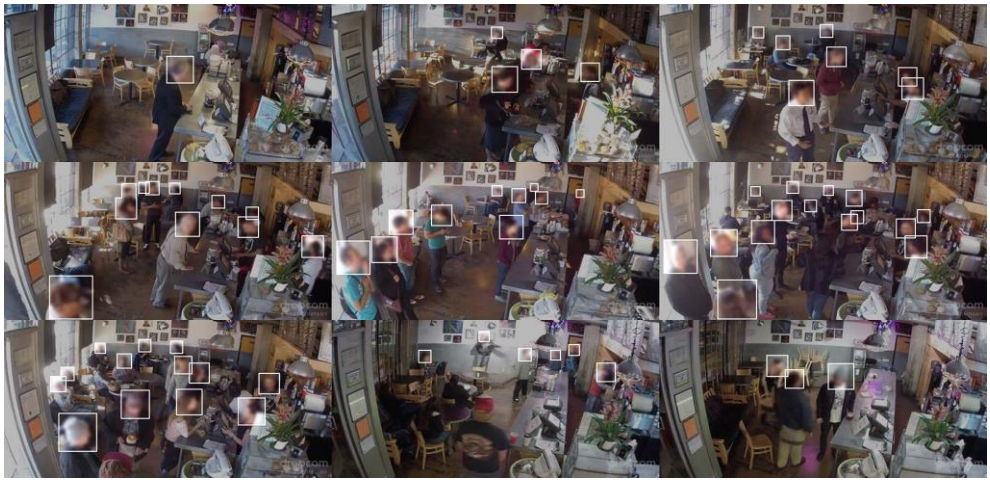

Fig. 2 Brainwash dataset

In this data set, 2000 pictures were selected as the training set, and the original coco model was used as the pre-training model, based on which training was performed to improve the detection performance of the system.

YOLO v3 uses the K-means clustering algorithm to cluster the data set to obtain the size of the anchor box, and the width and height of the target box are obtained through the width and height regression of the anchor box. The maximum number of bounding boxes that can be predicted by each grid of the output layer depends on the number of anchor boxes. It can be seen that the parameters of the anchor box have a significant impact on YOLO v3.

The original size of the nine anchor boxes set in YOLO v3 is based on 80 types of objects in the COCO dataset. In the COCO data set, the size of the target object is very different, and in the application scenario of this article, the proportion of objects in the picture changes little and the category is single, so the size of these anchor boxes is not completely applicable to the application scenarios proposed in this article[17]. Based on this, this paper uses the adjusted data set to re-cluster the anchor box to generate a size that is more in line with the characteristics of the scene, so as to improve the detection effect.

3. Result

Our test dataset includes pictures collected by ourselves and some data in Brainwash dataset, and we use mAP to evaluate the detection accuracy. Since only one category is involved in this system, the mAP value is exactly the value of AP, and the value of AP is the area under the P-R graph drawn by the recall rate and accuracy rate. Our experiment uses the Keras version of YOLO v3 algorithm, selects some pictures in the Brainwash dataset as the training dataset, and uses the size of the anchor box re-clustered according to the training dataset for training.

The experimental results show that the improved YOLO v3 algorithm has better performance in head detection, and the detection effect is significantly improved, which shows the effectiveness of the improved algorithm. The comparison of specific performance with other models is shown in Table 1 below.

Table 1 Comparison of test results before and after improvement

Model	mAP(%)
YOLO v3	19.82
YOLO v3(replace training set)	25.70
YOLO v3(replace training set + improve anchor box)	38.44

4. Conclusion

Fig. 3 test results

The experimental results show that, on the Brainwash dataset, the improved method has significantly improved detection accuracy compared to the original. Especially in the pictures taken by ourselves, the detection effect is better and can meet the expected goals. The test results of some test samples are shown in Figure 3 below.

Based on YOLO v3, our method improves the impact of problems such as target occlusion and small target objects on detection by replacing datasets and adjusting the size of the anchor box, and improves detection accuracy.

After testing, its accuracy can meet the needs of use and can be put into use, but there are cases where objects such as curtains and bags are mistakenly recognized as human heads. In the follow-up work, we will further improve and improve the solution proposed in this article in response to this problem and the feedback obtained from the usage.

References

[1] Zhu L, Wong KH: Human tracking and counting using the kinect range sensor based on adaboost and kalman filter, Advances in Visual Computing. p.582–591.

[2] Cai Z, Yu ZL, Liu H, Zhang K: Counting people in crowded scenes by video analyzing, 2014 IEEE 9th conference on industrial electronics and applications. p.1841– 1845.

[3] GAO C, LIU J, FENG Q, et al.: People-flow counting in complex environments by combining depth and color in- formation. Multimedia Tools and Applications, Vol.75(2016) No.15, p.9315-9331.

[4] Redmon J, Farhadi A. YOLOv3: An incremental improvement. 2018 IEEE Conf. on Computer Vision and Pattern Recognition (2018).2767-2773.

[5] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (Las Vegas, NV, USA. 2016). 779–788.

[6] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (Honolulu, HI, USA. 2016). 6517–6525.

[7] CHEN Z B, YE D Y, ZHU C X, LIAO J K: Target recognition method based on improved YOLOv3. Computer System Application, Vol.29(2020) No.01, p.49-58.

[8] THAN J: Research on an improved YOLOv3 target recognition algorithm (Ph.D, Huazhong University of Science and Technology, 2018),p.10.

[9] FANG Z L: Research on pedestrian detection technology of road traffic environment based on YOLOv3 (Ph.D, South China University of Technology, 2019),p.15.